AFRL-RI-RS-TR-2019-113

THE ECONOMICS OF CYBERSECURITY RESEARCH DATA SHARING

THE UNIVERSITY OF TULSA

MAY 2019

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th ABW, Wright-Patterson AFB Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2019-113 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
TODD CUSHMAN
Work Unit Manager

/ S /
JAMES S. PERRETTA
Deputy Chief, Information
Exploitation & Operations Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
MAY 2019	FINAL TECHNICAL REPORT	APR 2017 – DEC 2018

4. TITLE AND SUBTITLE

THE ECONOMICS OF CYBERSECURITY RESEARCH DATA SHARING

5a. CONTRACT NUMBER
FA8750-17-2-0148

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
N/A

6. AUTHOR(S)

Tyler Moore

5d. PROJECT NUMBER
DSHT

5e. TASK NUMBER
UL

5f. WORK UNIT NUMBER
SA

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The University of Tulsa
800 S Tucker Dr
Tulsa, OK 74104

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIGA
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2019-113

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. PA# 88ABW-2019-2139
Date Cleared: 02 May 2019

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This final technical report describes the result of a research project investigating economic issues involving the sharing of cybersecurity research data. It describes an effort to investigate data use and production in cybersecurity research publications from 2012-2016. Evidence is presented that researchers regularly use public data as input to research, but only rarely make created data publicly available. Additionally, it is shown that publications that do create datasets and make them publicly available are cited more often than those that do not. Additionally, utilization of the DHS IMPACT platform is investigated. Attributes of datasets are identified that are associated with greater demand from research consumers. Additionally, the value of sharing taken place on the platform is estimated to be approximately $663 million.

15. SUBJECT TERMS
Cybersecurity; data sharing; incentives; science of cybersecurity

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**TODD CUSHMAN**
U	U	U	UU	23	19b. TELEPHONE NUMBER (*Include area code*) **N/A**

TABLE OF CONTENTS

Figure **Page**

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

Figure **Page**

1.0 SUMMARY

Cybersecurity research and practice has become increasingly data-driven. Cybercrime indicators and other data can be used to better quantify risks. Data also helps inform proactive defenses by enabling others to learn from what has been targeted previously. Researchers have been using cybersecurity datasets as input to their own work as well as producing datasets as outputs to research for several years. Unfortunately, such data is not always shared with the broader research community, which makes replicating results difficult and developing new innovations using existing data infeasible.

Despite its importance, a number of challenges can stymie the sharing of cybersecurity data. Legal and privacy issues are frequently volunteered as an impediment to sharing. A more basic barrier is that sharing can be costly (in terms of time and dollars), yet the benefits of such sharing accrue to others. In economics terms, sharing data creates positive externalities, which tends to result in less of the good (in this case, cybersecurity datasets) being produced than is socially optimal. Finally, the incentives to share data are not always strong. Competitive concerns can inhibit firms who offer security products or services from sharing data with others. Researchers invest substantial effort in amassing reliable datasets and may prefer to mine the data themselves, fearing that others might take credit for discoveries based on data they have produced. Pioneering efforts have encouraged data sharing by researchers, principally DHS's IMPACT program. These have succeeded in reducing some barriers to sharing, notably legal ones.

This research project's goal was to empirically examine data usage and production by researchers in order to construct a better picture of the prospects for cybersecurity data sharing. This goal was achieved in two ways. First, the project systematically examined the published cybersecurity research literature to identify what data is being produced and consumed, how prevalent sharing is, and to seek out any positive incentives for researchers to share data. The project produced a taxonomy of cybersecurity datasets based upon inspecting 965 papers published between 2012 and 2016. Key findings include that three quarters of existing datasets used as input to research were publicly available, but less than one fifth of datasets created by researchers are publicly shared. Using a series of linear regressions, it is found that those researchers who do make public the datasets they create are rewarded with more citations to the associated papers. This suggests that an under-appreciated incentive exists for researchers to share their created datasets with the broader research community.

The second approach for examining data usage and production by researchers was to analyze usage data collected for IMPACT. Using linear regressions, factors are identified that affect the popularity of datasets. Over 2,000 written explanations of intended use are studied to identify patterns in how the datasets are used. Finally, a quantitative estimate of the financial value of sharing on the platform based on the costs of collection avoided by requesters.

2.0 INTRODUCTION

Data is an essential input to cybersecurity research. It takes many forms, from reports of compromised websites to network topologies, and from geolocations of backbone routers to traces of anonymous marketplaces peddling illegal goods. Whereas historically, the development of security-enabling technologies such as cryptography could be designed from mathematical foundations alone, today's security controls usually require data as input to the technology's design and to evaluate its effectiveness.

Unfortunately, accessing cybersecurity datasets is often difficult for researchers. Unlike many forms of data, security datasets are often seen as especially sensitive for several reasons. First, the data collected by one party could reflect poorly on others, e.g., if it indicated that a particular network had poor operational security. Second, the data might reveal to adversaries what is known about their activities, inadvertently assisting them in their criminal activities. Third, concerns over inadvertent sharing of private data preclude some organizations from sharing widely.

An economic perspective can be helpful in identifying the incentives and barriers to sharing research data, as well as quantifying the value of making such data available. This project’s primary scholarly contributions are to (1) measure the prevalence of cybersecurity data use and dissemination in research activities, (2) identify any positive incentives to share data by researchers, (3) empirically examine how researchers have utilized data made available through the IMPACT sharing platform, and (4) quantify the value of data shared through IMPACT by estimating the costs of data provision.

3.0 METHODS, ASSUMPTIONS AND PROCEDURES

3.1 Methodology for inspecting cybersecurity datasets in scholarly activities

In order to study how researchers use and produce datasets, we first selected suitable publication venues to examine. We started by selecting the top four computer security research conferences, ACM Conference on Computer and Communications Security (CCS), USENIX Security Symposium (USENIX), IEEE Symposium on Security and Privacy (S&P), and Network and Distributed System Security Symposium (NDSS). We complemented these with outlets that regularly publish data-intensive research: Internet Measurement Conference (IMC), International Conference on Financial Cryptography and Data Security (FC), and the Workshop on the Economics of Information Security (WEIS). Finally, we included relevant workshops associated with top conferences: the AI & Security Workshop at CCS, Cyber Security Experimentation and Test (CSET) Workshop at USENIX Security, and the Workshop on Bitcoin and Blockchain Research at FC (BITCOIN). We collected papers from these conferences from 2012 to 2016, inclusively. Focusing on these conferences makes the research efforts tractable, but this necessarily limits the scope of our findings.

We first downloaded all of the papers and collected their citation information. We used DBLP to get information on the papers and their linked URLs, and then crawled 2,037 papers from their corresponding websites. We then obtained citation information for all papers from Google Scholar.

We constructed a binary classifier to distinguish dataset-related papers and non-dataset related papers. Dataset papers are defined as those with at least one dataset that was used or created during the research. Non-dataset papers are papers that do not include a dataset as defined above. To build the machine learning model, we manually classified 391 papers into data (209) and non-data (182) papers. These 391 papers are randomly selected from the set of 2,037 papers, while ensuring sufficient coverage of all venues and years from 2012 to 2016. To construct features, we first extracted a list of “base form" (i.e., case and tense insensitive) words for each paper using the textblob Python package. We also filtered all stop words using NLTK's build-in list. From each paper's final word list, we built a word vocabulary from all papers and computed a TF-IDF vector for each paper. We then constructed several models using the sklearn Python package including Multinomial Naive Bayes, Bernoulli Naive Bayes, Gaussian Naive Bayes, SVM Confusion Matrix, and Random Forest. We used 10-fold cross validation to evaluate the models. Random Forest was the most accurate, but it still has 21.1% false positive rate and 17.2% false negative rate.

We used this model to classify all papers, 1,129 of which were predicted to include data. We inspected a random sample of 356 predicted data papers, confirming that 308 had data and 48 did not. Additionally, we inspected a sample of 218 predicted non-data papers that did not include references to dataset names identified in the training set. In total, we confirmed the presence or absence of data in 965 papers. 517 papers have data, with 209 coming from the training set and 308 confirmed from the test set. 448 papers do not have data, with 182 from training set and 266 confirmed from the test set. This labeled dataset, along with the R scripts to analyze the data, were produced as deliverables to this project and are available to the public at DOI:10.7910/DVN/4EPUIA. Moreover, the full methodology and results are described in a publication (Zheng et al. 2018).

3.2 Methodology for analyzing data from IMPACT

The operators of the IMPACT program have shared with us information on dataset requests, namely:

1. all requests for data made to the platform, from its inception in 2006 through September 30, 2018;
2. time when datasets are made available;
3. purpose requests in which the requester outlines its intended use in free-form text;
4. attributes of the dataset (e.g., provider, restrictions on use, time period of collection).

In total, 14 providers have made available 209 distinct datasets. 2,276 distinct requests for these datasets have been made. Additionally, the IMPACT team shared the results of email inquiries sent to all requesters in the summer of 2018 asking about whether and how the data was used. Furthermore, several data providers shared information on their costs.

The data was used in three ways to evaluate the value of cybersecurity datasets provided by IMPACT. First, we examined the factors that affect how frequently datasets are used. Many variables could influence a dataset's popularity among researchers, from restrictions placed on its commercial use to the type of data being shared. We empirically examined multiple factors using linear regressions.

Second, we examined the free-form text of the purpose requests to improve understanding of how the data is intended to be used. Upon examining all 2,276 of these reasons, a taxonomy was developed to encompass the various types of purposes researchers have for requesting this data. Six distinct categories were identified: technology evaluation, technology development, data analysis, operational defense, education and unspecified. Any individual reason may be classified into one or more of these categories.

Third, we utilized an alternative method to quantify the value of datasets by regarding value as the cost avoided by data consumers not having to collect data themselves. Fortunately, such data is readily available, as the IMPACT program pays data providers to share their data with requesters. Eight IMPACT performers shared detailed cost estimates for a number of categories such as personnel and equipment. Annual figures from 2012-17 were provided. The median annual cost was multiplied by the number of distinct requests to compute an estimate for the value created by the IMPACT platform.

We recognize that the metric's validity rests on a number of assumptions that may not hold in each circumstance. We assume each request is independent. We assume that the researchers experience no other sunk costs or utilize any existing resources when provisioning data. We assume that outside researchers would not have to expend resources gaining a sufficient technical understanding of the data collection requirements. We also assume that outside researchers would exercise the same level of care in collecting the data that the IMPACT performers do. Even if these assumptions do not hold universally across all requesters, the metric nonetheless provides a valuable estimate of what the “true” value might be. The full methodology and results are described in a publication (Moore et al. 2019).

4.0 RESULTS AND DISCUSSION

4.1 Census of cybersecurity research datasets

4.1.1 Taxonomy of cybersecurity research datasets. We were unable to find any existing taxonomy of cybersecurity datasets suited to our needs, so we constructed one based upon inspecting datasets in research publications. We iteratively constructed the categories by placing datasets identified in papers into prospective categories, modifying the groupings when uncertainties arose over which category the dataset should belong. Once the categories and subcategories stabilized, each dataset was manually labeled with a category, subcategory, whether the dataset already existed or was created by the authors, and whether the data was made publicly available. Four top-level categories were identified: attacker-related, defender artifacts, user & organizational characteristics, and macro-level Internet characteristics.

We classified any data that is already deemed malicious (e.g., scams, malware), or is used by attackers (vulnerabilities, cybercrime infrastructures) as *attacker-related*. There are four subcategories. **Attacks** contain information on attempts to harm digital assets perpetrated intentionally by malicious actors. **Vulnerabilities** contain information on weaknesses in digital assets that can be exploited by an attacker. **Exploits** contain information on how attacks may be perpetrated, but not when a particular system has been targeted by a malicious actor. Finally, **cybercrime activities** describe unlawful activities distinct from attacks, as well as information on the infrastructure and operations used by malicious actors to perpetrate attacks.

People and organizations construct defenses such as firewalls or secure configurations to block or prevent attacks. Frequently, as a consequence of constructing these defenses, data is generated (e.g., logs of blocked connection requests). These *defender artifacts* include configurations and alerts. **Alerts** contain outputs of defender artifacts, such as firewall logs or blackhole traffic. **Configurations** contain information about how defender artifacts are set up and configured (e.g., SSL certificate configurations).

Many datasets study *user and organizational characteristics*. **User activities** contain information about users' or organizations' online behavior, such as tweets. **User attitudes** contain information about opinions or attitudes towards an issue, often gleaned through surveys. **User attributes** contain information about the characteristics of users or organizations themselves (e.g., user profiles).

The Internet does not only consist of human activities but also many technical protocols and traffic, which we refer to as *macro-level Internet characteristics*. **Applications** contain information about Internet end products and services such as websites, Android apps, bitcoin, extensions, or code. **Network traces** are usually network traffic dumps that not only contain information regarding the application level, but also information about lower layers. Data usually comes from a benign resource, like an organization's internal network, but malicious traffic might be included. **Topology** datasets contain information about relationships between Internet components, such as routing between autonomous systems. **Benchmarks** contain information about measurements of Internet performance, such as upload/download speed or end-to-end network reliability. Finally, **adverse events** contain information on events that harm digital assets where malicious intent has not been established (e.g., outages caused by routing misconfigurations).

The dataset taxonomy just described has been incorporated into the categories used by the IMPACT program.

4.1.2. Empirical analysis of research datasets. Of the 965 research papers we labeled, 517 (55%) included data in some form. 229 papers (24% of the total) did not create a new dataset, but used an existing dataset as input to their research. Meanwhile 288 papers created datasets, either from scratch or by using other data as input. Just 61 papers, or 6% of those examined, created a dataset and made it publicly available.

Table 1 breaks down the datasets used in the papers, according to what was done and whether the underlying data was publicly available. Note that a single paper can use or create multiple datasets. In total, 902 datasets were created or used in the 517 papers. For created data, the distinction is made between primary and derivative data. Primary data is created solely by the authors, while derivative data leverages at least one existing dataset in order to make a new dataset. We can see that 76% of existing datasets used by researchers as input to their work are publicly available. This makes sense because public datasets are easier to access. Unfortunately, once the researchers create datasets, they are much less likely to return the favor by publishing their own datasets. 81-85% of created datasets described in research papers are not made publicly available. While not surprising, this stark difference highlights the opportunities missed by researchers failing to reciprocate by making their own data publicly available.

Table 1. Number and percentage of datasets made public, split by whether or not the dataset was created by researchers or already existed.

	Not Public		Public	
Dataset Type	#	%	#	%
Created Deriv.	89	85	16	15
Created Prim.	213	81	50	19
Existing	129	24	398	76

We now study the prevalence of different types of data used in cybersecurity research, according to the taxonomy outlined in Section 4.1.1. The leftmost numerical column in Table 2 shows the percentage breakdown of datasets across categories. Macro-level Internet characteristics comprise 47% of the total datasets encountered, with another 23% for user and organizational characteristics. Datasets related to attack (22%) and defense (8%) fill out the remainder.

The next column in Table 2 looks at how datasets in each subcategory are used. It reports the percentage of datasets in each subcategory that are created by the research, as opposed to re-use of existing data. Differences in proportion that are statistically under- and over-represented (according to a chi-squared test) are indicated. For example, we note that 71% of datasets describing vulnerabilities are created, compared to just 30% of datasets of attacks. This may indicate that attack datasets are particularly valuable inputs to research, or that vulnerabilities are more likely to be identified than subsequently used by others. Similarly, network traces, benchmarks and adverse events are disproportionately likely to be created rather than used. Note that low levels of reuse could also reflect difficulty sharing data, as is likely for network traces and user attitudes.

The last column in the table reports the proportion of datasets (created or existing) that are public. Exploits, application, topology and user attribute datasets are more likely to be public, while alerts, network traces, user activities and user attributes are less so.

Table 2. Incidents of datasets split by subcategory, plus proportion of datasets in each subcategory that are created and made public. Statistically significant under- and over-representations are indicated in bold with a (+/-).

	% Datasets	% Created		% Public	
Attacks	13	**30**	(–)	53	
Vulnerabilities	5	**71**	(+)	39	
Exploits	3	29		**75**	(+)
Cybercrime Inf.	1	56		44	
Alerts	3	30		**74**	(+)
Configurations	5	55		48	
Applications	24	36		**62**	(+)
Network Traces	9	**60**	(+)	**22**	(–)
Topology	9	**22**	(–)	**67**	(+)
Benchmarks	3	**81**	(+)	34	
Adverse Events	2	**67**	(+)	33	
User Activities	12	38		**41**	(+)
User Attitudes	1	**90**	(+)	**10**	(+)
User Attributes	10	**26**	(–)	**66**	(+)

4.1.3. Regression analysis of factors affecting citation rates. The preceding analysis has identified that while datasets are frequently created and used by cybersecurity researchers, it is uncommon for these datasets to be shared by the researchers who gather the data. There are many reasonable (and not so reasonable) explanations, including privacy considerations, restrictions on sharing by partners, and competitive concerns. On top of all that, it can be expensive and time-consuming to prepare the data for sharing and to accommodate requests.

Despite these downsides, there are definitely benefits to publishing datasets that accrue to the researcher. We explore one possible such benefit that is highly prized by academic researchers, namely, citations to the paper that created the dataset. We hypothesize that papers describing publicly available datasets will be cited more often, since other researchers can apply the data in their own follow-up work.

The summary statistics are encouraging. Papers that do not involve data or only use existing datasets in their work are cited 10 times per year (median), compared to 9.3 citations per year for papers that create datasets but don't publish them. By contrast, papers that do publish their data receive a median of 14.2 citations per year.

We constructed several linear regressions using the number of citations as the response variable. The explanatory variables include (i) # years since published, (ii) publication venue (using ACM CSS as baseline), (iii) a Boolean value indicating whether the paper created a public dataset, and (iv) dataset category (using attacks subcategory as baseline). Table 3 presents the results of four linear regressions that incrementally incorporate these explanatory variables. These regressions include only those papers that create datasets, in order to evaluate our key hypothesis that publishing datasets will be associated with higher citations. Similar results were found when including all papers.

The baseline model (1) finds that, as expected, time since publication affects citation rates. Each additional year since publication corresponds to 23 more citations. Approximately 10% of the variance in citation rates can be explained by time since publication alone. Adding in publication venue (model 2) explains a further 6.3% of the variance in citation rates. Papers creating datasets and published in the CSET, AISEC, and BITCOIN workshops are less likely to be cited than those in CCS, while those appearing in IEEE S\&P are considerably more likely to be cited than papers from CCS. Citations for other outlets (FC, IMC, NDSS, USENIX Security, and WEIS) were indistinguishable from CCS.

Table 3. Linear regression tables for papers that create datasets.

	Dependent variable:			
	citeNum			
	(1)	(2)	(3)	(4)
Years Published	23.059***	24.957***	25.619***	24.779***
FC		−26.982	−26.848	−24.712
IMC		−17.616	−23.730	−20.464
NDSS		−11.401	−15.367	−11.330
IEEE S&P		60.211***	55.741**	29.723**
USENIX Security		4.586	−0.717	−3.582
WEIS		−25.607	−27.932	−30.750
Workshops		−46.998**	−48.271**	−54.410***
Created Public			30.718**	24.651**
Vulnerabilities				−33.029*
Exploits				−29.843
Cybercrime Inf.				−2.050
Alerts				−51.072*
Configurations				−22.363
Applications				−12.232
Network Traces				−30.925*
Topology				−37.760*
Benchmarks				−36.534*
Adverse Events				−36.323
User Activities				−10.679
User Attitudes				−26.017
User Attributes				−14.081
Constant	−16.172	−16.412	−21.488	2.895
Observations	288	288	288	453
R^2	0.099	0.162	0.176	0.192
Adjusted R^2	0.096	0.138	0.149	0.151
Note:				*p<0.1; **p<0.05; ***p<0.01

Model 3 adds in a Boolean variable for whether the created dataset was made public. It is positive and statistically significant. The coefficient can be interpreted as papers that publish their datasets receive a boost of around 31 citations. Model 4 adds in dataset subcategories. Relative to attack datasets, papers that create datasets of vulnerabilities, alerts, network traces, topology and benchmarks are cited less often. Note that the number of observations is higher for model 4 because the unit of analysis is datasets in order to compare citations by dataset category.

We note that the boost to R^2 from models 3 and 4 is modest. Consequently, while we can conclude that making datasets public and the type of dataset created do meaningfully affect citation rates, there is much unexplained variance in citation rates beyond what is captured by these simple regression models. This is to be expected, since there are many characteristics of

papers that affect their likelihood to be cited that we do not consider (e.g, topical relevance, author reputation and influence, media attention).

4.2 Valuing cybersecurity dataset sharing on the IMPACT platform

We first examine how characteristics of datasets shared on IMPACT affect their popularity among requesters in Section 4.2.1. We then examine the stated purposes for requesting datasets in Section 4.2.2, and finally we construct an empirical estimate of the value of sharing data on IMPACT in Section 4.2.3.

4.2.1 Regression analysis of IMPACT dataset requests. The first way in which we evaluate the value of cybersecurity datasets provided by IMPACT is to examine the factors that affect how frequently they are used. Many variables could influence a dataset's popularity among researchers, from the restrictions placed on its commercial use to the type of data being shared. We empirically examine multiple factors using linear regressions with two distinct response variables: (1) the total number of requests a dataset receives and (2) the total number of approved requests. For these models, we only considered requests from 2016 onward because IMPACT utilization was relatively stable during this period. Explanatory variables include (i) time (in years) a dataset has been available to researchers since January 2016, (ii) dataset age (in years), (iii) a Boolean value indicating whether commercial use is allowed, (iv) restriction type (unrestricted, quasi-restricted, or restricted), (v) a Boolean value indicating whether data collection is ongoing, and (vi) dataset subcategory based on the taxonomy described in Section 4.1.1 using alerts as a baseline.

The tables in Table 1 present the results of the linear regressions. Surprisingly, the baseline model does not find the amount of time a dataset is available to researchers to significantly affect the number of requests it receives, though the overall age of the dataset is negatively correlated with requests. Adding in variables that cover access restrictions (model 2) yields more surprises. On their own, these variables have limited effect. None of the variables are significant for the regression measuring requests. Restricted datasets do receive fewer approved requests than do unrestricted datasets, however, and that difference is statistically significant. Furthermore, in Model 2, permitting commercial access does not affect utilization. However, the variables become significant and *negative* once additional explanatory variables are added in Model 3. In other words, permitting commercial use is associated with a reduction in requests. Additionally, quasi-restricted datasets are requested more often than unrestricted datasets, statistically significant at the 10% level. One possible explanation is that the more attractive datasets place more restrictions on access.

Model 2 alone explains roughly 3.7% and 8.3% of the variance in total requests and approved requests respectively. Adding in whether collection is ongoing and the dataset category (model 3) helps explain a lot more of the variance: 24% and 30% respectively. Ongoing collection corresponds to six more dataset requests. Topology and adverse event datasets are requested less often than alerts, while attacks are requested more often. In the request regression, configurations are also weakly underrepresented.

Table 4. Linear regression results for all requests (left) and approved requests (right).

	Dependent variable:		
	(Requests)		
	(1)	(2)	(3)
Constant	5.814**	6.339**	7.613*
Request Time	1.922	2.354*	3.528***
Age	−0.729***	−0.604**	−0.859***
Comm. Allowed		−3.357	−6.821**
Restricted		−0.379	−2.546
Quasi-Restricted		2.771	3.510*
Ongoing			6.607***
Configurations			−12.953*
Attacks			6.742**
Adverse Events			−7.589*
Applications			−5.031
Benchmark			−5.993
Network Traces			2.442
Topology			−5.610*
Observations	196	196	196
R^2	0.044	0.062	0.289
Adjusted R^2	0.034	0.037	0.238
Residual Std. Error	10.224 (df = 193)	10.209 (df = 190)	9.082 (df = 182)
Note:			*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:		
	(Approved)		
	(1)	(2)	(3)
Constant	4.640**	5.584**	5.915*
Request Time	1.524	1.929*	3.002***
Age	−0.653***	−0.535**	−0.748***
Comm. Allowed		−2.385	−4.885**
Restricted		−3.204**	−5.269***
Quasi-Restricted		1.832	2.369
Ongoing			5.054***
Configurations			−9.424
Attacks			6.203**
Adverse Events			−6.538*
Applications			−2.698
Benchmark			−4.253
Network Traces			2.615
Topology			−4.536*
Observations	196	196	196
R^2	0.053	0.107	0.342
Adjusted R^2	0.043	0.083	0.295
Residual Std. Error	8.505 (df = 193)	8.325 (df = 190)	7.302 (df = 182)
Note:			*p<0.1; **p<0.05; ***p<0.01

4.2.2 Empirical analysis of dataset usage. Valuing information goods such as cybersecurity datasets is fraught with difficulty. The most obvious approach is to assign a value corresponding to the amount others are willing to pay to obtain it. This is not an option for public goods like IMPACT datasets that are given away for free. An alternative is to investigate how others use the data, thereby creating value. This is a worthwhile approach because it can shed light on the outputs or outcomes that result from data use. The challenge with this approach is that it is hard to aggregate the myriad uses into a single dollar estimate of value. We defer until Section 4.2.3 a discussion of a method to provide a dollar estimate of IMPACT datasets.

Six categories are identified. Requests are categorized as **technology evaluation** when datasets are used to evaluate the effectiveness of some technology. This may be an algorithm, framework, model, application, theory or any other form of technology that the requester wishes to test. An example technology request is “need to evaluate if our new DDoS detection in-line analytical module in NetFlow Optimizer can detect this attack.”

Technology development requests utilize data to assist with the development of some technology. The requester may wish to extract features from the dataset that aid them in developing a technology. Datasets that are used to train machine learning applications are also considered technology development. An example request is to “incorporate the attack scenarios to devise an automated process of detecting and controlling malicious insiders to mitigate risks to the organization."

Under **data analysis**, the requester wishes to analyze the data for its own sake. Data analytics, data visualization, and characteristic extraction all fall under data analysis. An example request is “the data will be used to analyze how DDoS affects the open source production systems."

For **operational defense** requests, data helps protect some critical resource of the requester's organization. Requesters may want to see if the data reports on resources controlled by their organization or if the data can help strengthen their defenses. An example request reads “my objective is to protect Marine Corps data. This database can provide intelligence on passive DNS malware that can be used to block it from entering my network”.

Data can also be requested for **education** purposes such as use in courses or clubs in a school setting such as a university or high school. An example request is “I'd like to develop exercises for an introductory stats and data science course that emphasizes cybersecurity awareness for the state of Virginia."

Finally, some requests remained **unspecified**. In this case, the request reason was either too vague or unclear. Examples include “Need for research”, and “I'm doing some research on cyber situation awareness and feel this data would be beneficial to this work”.

We manually categorized each request according to this taxonomy described above. Requests could correspond to more than one category, or to no category at all. Data analysis was most common (31%), followed by 28% each for technology evaluation and development. 6% of requests fell into the operational support category, with 3% in education.

We additionally sought to understand not only what the dataset was requested for, but also what it was ultimately used for. DHS surveyed all IMPACT requesters whose requests had been approved, receiving 114 responses. When asked whether or not they actually used the dataset they had requested, 60% of respondents said they had. To better understand what those requesters actually used the dataset for, we asked them to categorize their request reason and to categorize what their actual use was using the request taxonomy described above. 91% of requesters reported that they used the datasets in the same manner that they originally requested. This suggests that the preceding analysis on intended use accurately reflects actual use.

Furthermore, we asked the requesters who used the dataset whether or not they would have collected the themselves had IMPACT not provided the dataset. 72% answered that they would not have collected the data themselves. For those that wouldn't have collected the data themselves, their research may not have continued. For the 28% that would have collected the data themselves, they would have been replicating costly data collection and wasting time or resources that could be spent elsewhere. Fittingly, the next subsection describes a quantitative model of value based on the avoided cost of data collection.

4.2.3 Quantifying value through avoided cost of data collection. Table 5 reports the median cost figures for each category reported by IMPACT performers, along with the total of approximately $291K. Given that IMPACT has shared data with 2,276 requesters, the total value created as measured by this metric since the program's inception in 2006 is $663 million.

Table 5. Median reported annual cost of providing datasets to IMPACT.

Category	Cost
# Personnel	3
PI	$38,500
Software Developer	$87,000
System Administrator	$80,000
Research Staff	$30,825
Managerial Cost	$37,000
Equipment	$18,250
Total	**$291,575**

Figure 1 plots the annual value created for all requests (solid red line), as well as a more conservative measure that normalizes for intended use (green dashed line). A normalization factor of 60% is used since that is the proportion of surveyed recipients who reported using the dataset they requested.

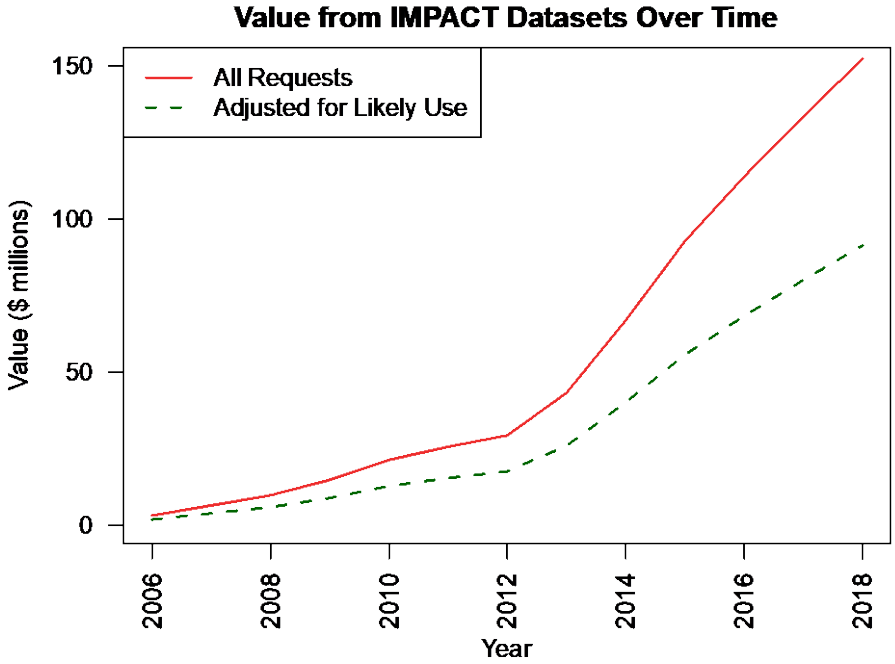

Figure 1. Value of data shared by IMPACT since inception using avoided cost definition.

Figure 2 splits the value created among the top four IMPACT providers who have shared annual costs. We can see considerable variation, which is a consequence of highly variable costs of data production and dataset popularity.

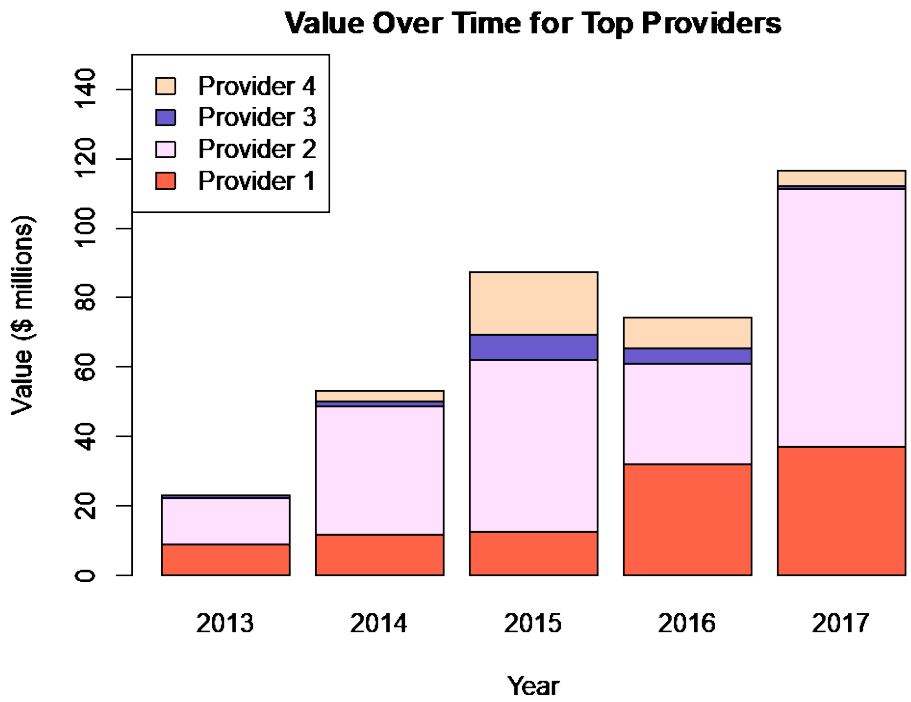

Figure 2. Value of data shared by top 4 providers on IMPACT using costs reported for the year in which data was requested.

We next investigate the relationship between the cost of producing a dataset and its ensuing demand from users. Table 6 lists the cost per dataset request for each of the top providers. We can see that the cost per request varies by an order of magnitude.

Table 6. Cost per request among top 4 providers.

Provider	Cost Per Request
P1	$13 394
P2	$10 056
P3	$3 507
P4	$1 567

Digging deeper, we observe little to no relationship between the number of requests a dataset receives and what it costs to produce. Figure 3 plots the annual provider cost against the number of requests received that year for the top 4 providers. The best-fit line indicates a very slight positive correlation between cost and requests, but it is clear that many other latent factors besides cost of production affect a dataset's popularity.

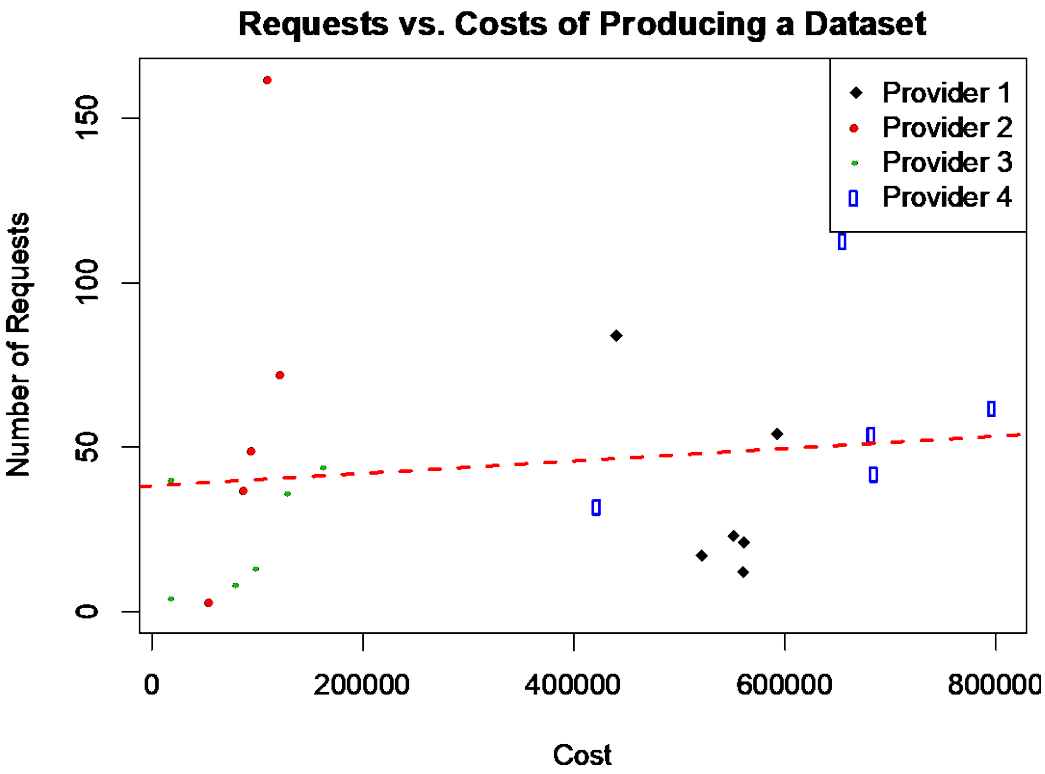

Figure 3. Scatter plot of provider annual requests compared to cost of providing data.

5.0 CONCLUSIONS

This research project has significantly advanced knowledge in two areas related to the economics of cybersecurity research datasets. First, it has empirically examined the use and production of research data in the academic literature on cybersecurity. Second, it has studied usage of the pioneering IMPACT platform and quantified the value it has created.

For the first area of investigation, the project has conducted a first-of-its-kind examination of cybersecurity research publications to identify what datasets are created, how they are used, and how to encourage more sharing among researchers. By examining nearly 1,000 papers, we have taken a data-driven approach to constructing a taxonomy of cybersecurity research datasets, which then sheds light on which types of data are being created, used and shared with the broader community. Some findings underscore the disincentives to make datasets publicly available to others. While 76% of existing datasets used by researchers are publicly available, just 18% of created datasets return the favor. Despite increasing attention being paid to cybersecurity research involving data and exhortations to share data publicly, the proportion of datasets that are shared publicly has remained consistently low.

So what can be done to disrupt the status quo? Paying attention to the incentives to share by eliminating barriers and rewarding publication is key. This project has made a start in that direction by identifying that citation rates are higher for papers that make created datasets publicly available. To the extent that this finding and subsequent research can shift the narrative about data sharing away from community service towards it being individually rational, we believe more researchers will elect to publish datasets and the science of security can be advanced.

For the second area, we have empirically investigated the sharing that has taken place on IMPACT, a long-running platform that has uniquely facilitated free access to cybersecurity research data. Controlling for the time available on IMPACT, we have found that the dataset's age is negatively correlated with requests. This makes sense given that researchers may prefer more recent data for their efforts. We also found that the restrictions placed on access to data affect how often they are requested, but in unexpected ways. For example, permitting commercial use of the data is negatively correlated with utilization, and quasi-restricted datasets are requested more often than unrestricted ones. These may reflect either a perception (or the reality) that datasets placing modest restrictions are more likely to be useful. Note that when we do move to the restricted category that introduces significant additional costs and verification, approved requests fall. We also found that datasets that are made available on an ongoing basis are requested more often. Ongoing availability can be thought of as a proxy for current relevance and longitudinal cohesiveness, two properties valued by researchers.

We also found that there is considerable variation among the types of datasets. Twenty percent of the variance in requests can be explained by the type of data offered and whether or not it is made available on an ongoing basis. Difficult to collect, topically relevant, and potentially sensitive data such as attacks are requested more often, while more general and less sensitive data such as network topology are requested less often.

We also investigated the value created by data shared on IMPACT in two ways. First, we looked at what the requesters themselves said they intended to do with the data. We identified five categories of use: technology evaluation, technology development, data analysis, operational defense, and education. Data analysis was the most common intended use, followed by

technology development and evaluation. Strikingly, when asked, 60% of requesters said they used the data requested and 90% of those said they used it in the way they originally intended. This suggests that the IMPACT users are highly sophisticated in their understanding of their research data needs. Most significantly, 72% of surveyed requesters stated that they would not have collected the datasets themselves if they could not have obtained it through IMPACT. This highlights the value of investing in research data infrastructure and underscores how much research may not be conducted when data access is limited.

This motivates the second approach to valuing data shared on IMPACT, by quantifying value in terms of the costs avoided by data recipients. We obtained annual provisioning costs from data providers. Matching this to requests, we estimate that the value created since program inception in 2006 is $663 million. Digging deeper into the costs uncovers two surprising insights. First, the normalized cost per request varies widely, by one order of magnitude. Second, there is little if any relationship between the cost of data provisioning and its resulting demand.

On one level, it is not surprising that the relationship between the cost of data production and the resulting demand for it is weak at best. What drives researcher interest is how the data can be leveraged, not the person-hours required to collect the data in the first place. Nonetheless, the implications for funding cybersecurity research data production are significant. Ideally, program managers should (and assuredly do) consider the potential demand for a dataset when deciding whether to support an effort financially. But perhaps more weight should be given to the anticipated requests per unit cost in order to maximize the impact of limited budgetary resources. To do so would also require more work estimating the demand of datasets in advance. To an extent, the regressions described in Section 4.2.1 can help identify dataset categories that are in higher demand, but more work is needed to test whether such retrospective analysis is predictive of future demand.

6.0 REFERENCES

T. Moore, E. Kenneally, M. Collett, and P. Thapa, “Valuing cybersecurity research datasets”, in submission.

M. Zheng, H. Robbins, Z. Chai, P. Thapa, and T. Moore, “Cybersecurity research datasets: Taxonomy and empirical analysis,” *11th USENIX Workshop on Cyber Security Experimentation and Test (CSET 18)*, Baltimore, MD, 2018.

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

ACM	Association for Computing Machinery
AISEC	Artificial Intelligence and Security
CCS	Computer and Communications Security
CSET	Cyber Security Experimentation and Test
DBLP	Computer Science Bibliography
DHS	Department of Homeland Security
IEEE S&P	Institute for Electronics and Electrical Engineers Symposium on Security and Privacy
IMC	Internet Measurement Conference
IMPACT	Information Marketplace for Policy and Analysis of Cyber-Risk and Trust
NDSS	Network and Distributed Security Symposium
NLTK	Natural Language Toolkit
PI	Principal Investigator
SSL	Secure Socket Layer
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
URL	Uniform Resource Locator
WEIS	Workshop on the Economics of Information Security